

# SIPFormer: Segmentation of Multiocular Biometric Traits With Transformers

Bilal Hassan<sup>id</sup>, Taimur Hassan<sup>id</sup>, *Member, IEEE*, Ramsha Ahmed<sup>id</sup>, Naoufel Werghi<sup>id</sup>, *Senior Member, IEEE*, and Jorge Dias<sup>id</sup>, *Senior Member, IEEE*

**Abstract**—The advancements in machine vision have opened up new avenues for implementing multimodal biometric identification systems for real-world applications. These systems can address the shortcomings of unimodal biometric systems, which are susceptible to spoofing, noise, nonuniversality, and intra-class variations. Besides, ocular traits among various biometric traits are preferably used in these recognition systems due to their great uniqueness, permanence, and performance. However, segmenting visual biometric features under unconstrained situations remains challenging due to a variety of variables, such as Purkinje reflexes, specular reflections, eye gaze, off-angle pictures, poor resolution, and numerous occlusions. To overcome these challenges, this research presents a novel framework called SIPFormer, comprising the encoder, decoder, and transformer blocks to simultaneously segment three ocular traits (sclera, iris, and pupil) using its discriminative multihead self-attention mechanism. Besides, we used the large publicly available iris database reflecting different unconstrained acquisition settings, with inherent noise effects such as scanner artifacts, intensity and illumination variations, motion blur, and occultations caused by eyelashes, eyelids, and eyeglasses. Furthermore, the simulation results demonstrate the efficacy of the proposed SIPFormer model, where it achieved the mean Dice similarity coefficient scores of 0.9018, 0.9176, and 0.9229 for segmenting the sclera, iris, and pupil classes, respectively.

**Index Terms**—Biometric traits, iris, pupil, sclera, segmentation, transformers.

## I. INTRODUCTION

OVER the last decade, the need for reliable authentication systems has grown in step with the meteoric rise of the information technology industry and rapid technological development. As a result, researchers are constantly striving

Manuscript received 26 July 2022; revised 26 October 2022; accepted 28 November 2022. Date of publication 26 December 2022; date of current version 13 January 2023. This work was supported by Khalifa University Center for Autonomous Robotic Systems (KUCARS). The Associate Editor coordinating the review process was Dr. Hongrui Wang. (*Corresponding author: Taimur Hassan.*)

Bilal Hassan and Jorge Dias are with the Khalifa University Center for Autonomous Robotic Systems (KUCARS), Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates (e-mail: bilal.hassan@ku.ac.ae; jorge.dias@ku.ac.ae).

Taimur Hassan and Naoufel Werghi are with the Khalifa University Center for Autonomous Robotic Systems (KUCARS) and the Center for Cyber-Physical Systems (C2PS), Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates (e-mail: taimur.hassan@ku.ac.ae; naoufel.werghi@ku.ac.ae).

Ramsha Ahmed is with the Healthcare Engineering Innovation Center (HEIC), Department of Biomedical Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates (e-mail: ramsha.ahmed@ku.ac.ae).

Digital Object Identifier 10.1109/TIM.2022.3232162

to perfect foolproof authentication methods. Biometric technology, which refers to the authentication of a person based on measurable physical or behavioral attributes, is becoming increasingly popular in this area. Moreover, biometric systems have a remarkable false match and false rejection rate of 2%, making them almost hard to exploit using standard decryption approaches [1]. To a large extent, biometric systems are now required in our daily lives. Unlike conventional methods, these alternatives do not need us to physically store or remember sensitive information, such as user names, passwords, or other authentication credentials. Biometrics are used in various crucial applications, from unlocking mobile phones to cash withdrawal, and consumer apps to law enforcement and restricted access control [2], [3], [4].

Several biometric identifiers can be used to identify an individual positively. Among these, ocular features have proven superior to other biometric attributes for applications requiring high reliability and accuracy due to their dependability, longevity, and efficiency [5]. In contrast, systems that rely on other characteristics, such as fingerprints, can be easily compromised, as they may be burned or affected by allergic skin reactions with time. Similarly, the performance of the voice recognition system is unreliable since voices can be manipulated [6], [7]. The primary ocular biometric traits are the sclera, iris, and pupil, as shown in Fig. 1. Each ocular trait has its own uniqueness and importance, as described in the following.

- 1) *Sclera*: A relatively new biometric trait for person identification has shown promising results [8], [9]. The vascular pattern in the sclera (see Fig. 1) is highly unique for each individual and even observed to be different between the left and right eyes of a person [10]. In addition, it is tough to counterfeit the sclera, unlike the iris, which can be easily forged by wearing a contact lens [10]. Moreover, segmenting the sclera can help achieve higher accuracy of iris recognition systems under unconstrained lighting conditions [11].
- 2) *Iris*: The most widely used ocular trait in biometric systems possesses a high degree of distinctiveness and randomness in terms of its pattern, size, shape, and color. This complexity is primarily because of the rich and unique textures of the iris, such as furrows, rings, freckles, crypts, zigzags, or ridges [12], as shown in Fig. 1. Moreover, the iris trait exhibits greater immutability

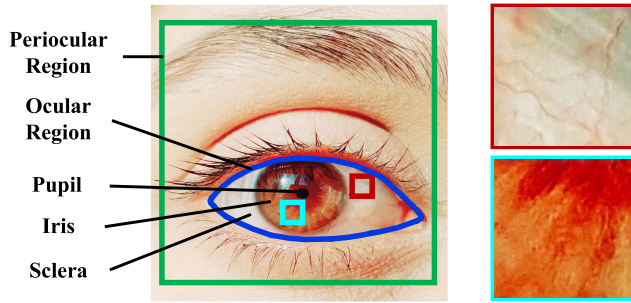


Fig. 1. Periocular and ocular components. The red and aqua boxes represent the vascular pattern of sclera and iris texture, respectively.

throughout a person's life, and some studies even concluded the usefulness of the iris in postmortem recognition [13].

- 3) *Pupil*: The least commonly used ocular trait in biometry due to its homogenous structure. Generally, segmentation and detection of the pupil are considered the fundamental procedure in developing various computer vision applications [14]. However, there are limited studies based on the pupil as a standalone trait for authentication [15].

The characteristics of these ocular traits in conjunction vary extensively across the human population. Over time, uniqueness, and randomness, their immutability can provide a robust and reliable multimodal biometric recognition system. Despite these advantages, joint segmentation of ocular traits is a challenging task mainly due to three major factors: Purkinje reflexes, eye gaze, and occlusions due to eyelids and eyelashes.

Given the above, the motivation of this work is to present a robust framework that can jointly segment the three ocular traits (sclera, iris, and pupil), facilitating the development of a biometric system based on multiocular traits in the future. A recent spurt in the expansion of deep learning applications may be attributed to the proven effectiveness of various convolutional neural network (CNN) architectures over other traditional approaches [16], [17], [18], [19]. These advantages have proliferated deep learning-based biometric systems, and the domain has surged recently in security and authentication applications with a significant emphasis on ocular traits [20].

#### A. Related Works

- 1) *Sclera*: The researchers have previously proposed different deep learning solutions to segment and recognize the sclera for biometric applications [10], [11], [21], [22], [23]. Maheshan et al. [10] proposed a CNN sclera recognition engine consisting of four convolutional units and one fully connected unit. They evaluated the framework on the sclera segmentation, and recognition benchmarking competition dataset and received an accuracy of 87.65%. Besides, the method for semantic segmentation of sclera in [11] used the combination of CNN and conditional random fields (CRFs) as a postprocessing technique. They validated the framework on the sclera competition dataset and received an accuracy of 83.2% in the correct classification of sclera pixels.

Moreover, Zhu et al. [21] designed a stem-and-leaf branches network, called SLBNet, to identify persons. They first used the traditional image processing techniques to segment the scleral vasculature, which is then passed on to the SLBNet to identify the person. Similarly, different neural network architectures are implemented in [22] to segment the iris and sclera using two different datasets. Furthermore, in [23], a CNN model called ScleraNET is presented to identify and recognize a person using a sclera vasculature pattern.

- 2) *Iris*: Most of the existing works related to ocular biometry in the literature have been conducted using the iris trait [20]. In the past, many researchers proposed different iris recognition methods. The most common ones include feature descriptor-based methods [24], [25], [26]. Recently, researchers have implemented deep learning-based iris segmentation and recognition frameworks. Jha et al. [27] proposed an iris segmentation framework at the pixel level (PixlSegNet). Their framework is based on the convolutional encoder–decoder architecture, where a stacked hourglass network is used between the encoder and decoder paths. Besides, Nguyen et al. [28] evaluated the performance of six different pretrained CNN architectures on iris recognition using two publicly available datasets. They showed that standard CNN features, originally extracted and trained for classifying common objects, can also be transferred and used to recognize iris.

Moreover, the capsule network-based deep learning framework for the recognition of iris is proposed in [29]. Their algorithm adjusted the network structure detail to adapt for iris recognition based on a modified dynamic routing algorithm within the capsule layers. They employed the transfer learning approach and divided the three pretrained CNN architectures into subnetwork sequences to extract the features. Furthermore, the deep multimodal biometric system based on iris recognition is proposed in [5]. They first localized the iris regions in both the left and right eyes of the same person and then passed to the CNN for extraction of discriminative features and classification using the rank fusion technique. In addition to that, a fully convolutional deep neural network framework to segment iris using low-quality images is proposed in [4]. They merged four different CNNs using semiparallel deep neural network techniques.

- 3) *Pupil*: In the past, various deep learning solutions have been proposed to detect, segment, and track pupil. Yiu et al. [30] used a U-Net-based CNN architecture called DeepVOG to segment the pupil. They trained the network on two local datasets containing video-oculography (VOG) images. They validated the framework on different datasets and achieved the highest median value of the Dice coefficient as 0.978. Moreover, the approach in [31] is validated on two different datasets, consisting of a close-up view of the eye and the full facial image. In the case of the full image, the authors first extracted the eye region, which is then passed to the pupil segmentation network. In [32], another deep CNN called DeepEye is proposed for pupil detection based on atrous convolutions and spatial pyramids.

Furthermore, Whang et al. [33] used a lightweight CNN architecture to segment the pupil from the video sequences. They predicted the size of the pupil using the major and

minor axes of an ellipse. Besides, Shi et al. [34] proposed an end-to-end deep learning framework for pupil detection and tracking. They used a CNN approach to detect the pupil and the long short-term memory (LSTM) model to predict pupil motion. Similarly, Ou et al. [35] employed the pretrained deep learning-based detector for pupil-center detection and tracking in the visible-light mode.

### B. Contributions

The existing research for ocular biometrics is typically based on a single (mostly iris) or two ocular traits (mostly iris and pupil). The authors have implemented different deep learning-based algorithms to detect, segment, and recognize these ocular traits. Compared to previous methods, the proposed convolutional transformer-based framework, SIPFormer, can simultaneously segment all three ocular biometric features (sclera, iris, and pupil) and will also improve the accuracy of multimodal biometric systems for unconstrained conditions. The notable contributions of this research are twofold, as summarized in the following.

- 1) This article presents a novel end-to-end deep learning model called SIPFormer, comprising an encoder, decoder, and transformer blocks to perform joint segmentation of multiple ocular traits (sclera, iris, and pupil). Besides, to the best of our knowledge, the SIPFormer framework is the first attempt to utilize transformers with convolutional blocks to perform joint ocular traits segmentation.
- 2) Moreover, SIPFormer, due to its discriminative multi-head self-attention, possesses the intrinsic capacity to segment multiocular traits irrespective of the scanner artifacts and noisy shadows produced due to the presence of eyelashes, eyelids, and spectacles. The SIPFormer has been rigorously tested on five diversified datasets, demonstrating comparable segmentation performance with 136.21% fewer parameters than state-of-the-art deep learning-based segmentation methods.

The remaining of this article is organized as follows. Section II gives the description of the dataset used in this research and the proposed framework in detail. Next, Section III presents the experimental setup. Furthermore, the simulation results are presented in Section IV, followed by a discussion and conclusion in Section V.

## II. MATERIALS AND METHODS

### A. Dataset Details

In this study, we have opted for the Chinese Academy of Sciences, Institute of Automation (CASIA) database [36] mainly because it is one of the largest datasets with a lot more subjects and intraclass variations and our prime focus is toward the segmentation and extraction of ocular modalities for which this database is well suited. In this study, we used a total of 52034 images of about 2800 subjects for training and validation purposes. These images are retrieved from five subsets of the CASIA-IrisV4 database, which are CASIA-Iris-Interval (CII), CASIA-Iris-Syn (CIS), CASIA-Iris-Lamp

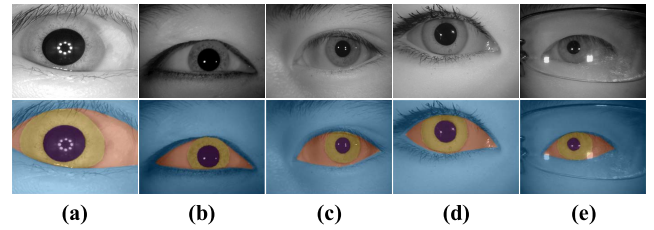


Fig. 2. Sample images from CASIA-IrisV4 subsets and corresponding ground-truth labels. (a) CII. (b) CIS. (c) CIL. (d) CITW. (e) CIT.

(CIL), CASIA-Iris-Thousand (CIT), and CASIA-Iris-Twins (CITW). The statistics and features of each of these five subsets are shown in Table I. Fig. 2 shows the sample images and the corresponding ground-truth labels.

We randomly divided each of the five subsets of the CASIA-IrisV4 database in the ratio of 60:20:20 for training, validation, and testing purposes, as shown in Table II. The proposed SIPFormer model is trained using 60% of the images in each subset, while for both validation and testing, we used 20% of the images in each subset.

### B. Data Preprocessing

In our study, we have used five different subsets of the CASIA-IrisV4 database that vary extensively in terms of environment, illumination, and camera sensor, as shown in Table I. Hence, before feeding to the CNN architecture, data preprocessing is required to scale, convert, and standardize these images according to the dimension, activation shape, and size specified by the network input specifications. Therefore, preprocessing is the first stage of the proposed SIPFormer system, where we first performed the intensity transformation to adjust and increase the intensity differences in the ocular region, as shown in Fig. 3(b). For this purpose, we employed two intensity transformation functions, gamma transformation, and contrast stretching, to pick out the details, such as limbus (iris–sclera boundary) in the ocular region.

Next, we enhanced the images using the local enhancement technique. First, we improved the dynamic range of the images using histogram equalization to evenly distribute the pixel intensities across the entire range. The histogram equalization process improved the contrast level, especially in images where the intensities are clustered predominantly around the lower or middle range. We further enhanced these images using the contrast-limited adaptive histogram equalization technique, as shown in Fig. 3(c).

After enhancing the images, we removed two types of reflections from the images: reflection in the ocular region mainly due to cornea and aqueous humor and reflection in the periocular region mainly due to flashlights. We used the adaptive thresholding scheme to identify the bright white spots in the images and filled regions using the morphological reconstruction technique. Compared to the conventional algorithms based on the global threshold value, the adaptive thresholding scheme utilizes the dynamic threshold value for each pixel in the image, computed using the local mean intensity in the pixel neighborhood. Finally, we resized the preprocessed images to a common resolution of  $576 \times 768$ . Fig. 3(d) shows the

TABLE I  
SPECIFICATIONS OF CASIA-IRISV4 DATABASE

Subsets	Images	Subjects	Resolution	Environment	Sensor	Features
CII	2,639	249	320×280	Indoor	CASIA NIR LED	Cross-session extremely clear iris images, detailed iris texture features
CIS	10,000	1,000	640×480	-	-	Synthesized iris images
CIL	16,212	411	640×480	Indoor	OKI IRISPASS-h	More intra-class variations, elastic deformation of iris texture, non-linear iris normalization
CITW	3,183	200	640×480	Outdoor	OKI IRISPASS-h	Iris images of 100 pairs of twins
CIT	20,000	1,000	640×480	Indoor	Irisking IKEMB-100	High-quality iris image, intra-class variations due to specular reflections

CASIA-IrisV4 is the latest version of the CASIA Iris Image Database, released to the international biometrics community in 2010 [36].

TABLE II  
DETAILS OF TRAINING, VALIDATION, AND TEST SETS

Subsets	Total Images	Training (60%)	Validation (20%)	Test (20%)
CII	2,639	1,583	528	528
CIS	10,000	6,000	2,000	2,000
CIL	16,212	9,728	3,242	3,242
CITW	3,183	1,909	637	637
CIT	20,000	12,000	4,000	4,000
Total	52,034	31,220	10,407	10,407

CASIA-IrisV4 is the latest version of the CASIA Iris Image Database, released to the international biometrics community in 2010 [36].

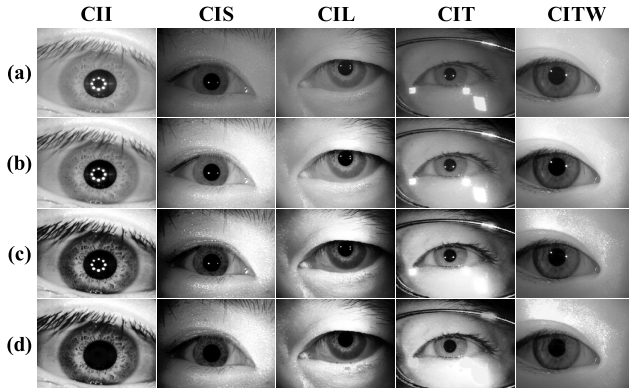


Fig. 3. Data preprocessing. (a) Pristine images. (b) Intensity transformation. (c) Image enhancement. (d) Reflection removal.

preprocessing stage results on randomly selected images from all five CASIA-IrisV4 subsets.

### C. Proposed SIPFormer Architecture

The proposed SIPFormer model is designed to perform the joint segmentation of the sclera, iris, and pupil from the periocular scans. The high-level overview of the proposed SIPFormer model is shown in 4. As evident from Fig. 4, the SIPFormer architecture consists of three units dubbed: the SIPFormer encoder, SIPFormer decoder, and the SIPFormer transformer. When the input scan is preprocessed, it is passed through the SIPFormer encoder, which generates the latent feature representations to distinguish the multiocular traits. Moreover, the SIPFormer encoder also serves as a backbone to generate latent projections from the nonoverlapping sequenced patches obtained from the candidate scan. Afterward, the latent projection and the flattened projections of the positional embeddings (generated through normalized cross correlation) are added and passed to the three-layered SIPFormer transformer. The SIPFormer transformer computes

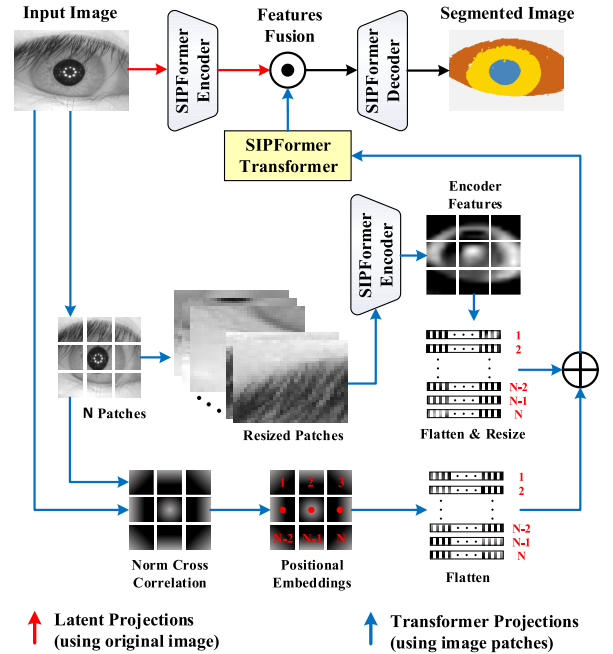


Fig. 4. High-level overview of the SIPFormer model.

the contextual multihead self-attention distribution from the scan projections, enabling the SIPFormer encoder to amplify the discrimination of ocular traits through the fusion between the SIPFormer encoder and transformer features via convolution. The resultant latent space distribution is passed to the SIPFormer decoder that generates the segmented scan with multiocular trait representations. Moreover, Fig. 5 shows the detailed SIPFormer architecture with layerwise configuration and connections. The detailed description of each unit within the proposed SIPFormer architecture is presented in the following.

1) *SIPFormer Encoder*: The SIPFormer encoder is responsible for generating the distribution of the latent feature  $f_e(x)$  to extract the multiocular traits from the periocular scans  $x \in \mathbb{R}^{(R \times C \times C_h)}$ , where  $\mathbb{R}$  denotes the rows,  $C$  denotes the columns, and  $C_h$  denotes the channels of  $x$ . Unlike the conventional pre-trained networks, the SIPFormer encoder consists of multiple trait preservation (TP) and residual (Res) blocks, as shown in Fig. 5. These blocks enable the SIPFormer encoder to produce an accurate contextual and semantic representation of the ocular traits during the scan decomposition to yield distinct feature maps. In total, there are four TP blocks and 12 Res blocks within the SIPFormer encoder, where

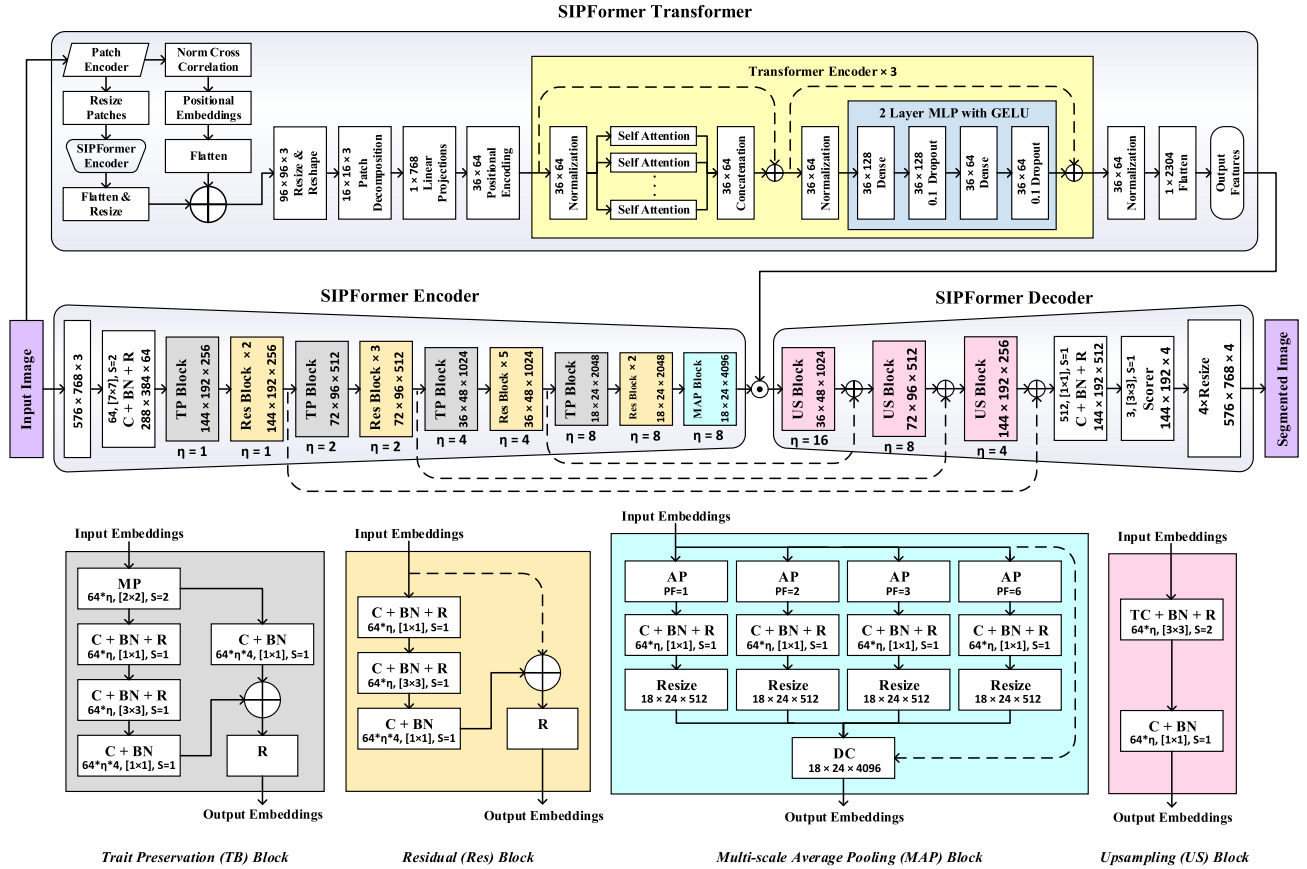


Fig. 5. Architectural details of the proposed SIPFormer model.

each TP block consists of four convolutions (C), four batch normalizations (BNs), two rectified linear units (R), and one max pooling (MP), whereas each Res block contains three C, three BNs, and three R's along with a skip connection. The latent features  $f_e$  produced by the SIPFormer encoder (after tuning its learnable weights) effectively discriminate the ocular trait representations from the noisy eyelashes regions and the background skin. However, it also results in false positives toward differentiating between sclera and iris regions as their features are very well correlated. To overcome this, we boost the separation of interclass distributions by convolving  $f_e$  with the transformer projections  $p_t$ , which yields the fused feature representation  $f_d = f_e * p_t$ . These fused features are passed to the SIPFormer decoder to extract ocular traits.

2) *SIPFormer Transformer*: As mentioned above, after generating  $f_e$  from the SIPFormer encoder, we fuse them with the transformer projections  $p_t$ , generated by the SIPFormer transformer unit to increase the interclass separability between the well-correlated trait distributions. The SIPFormer transformer consists of three encoders that are coupled together in a cascaded fashion to generate transformer projections  $p_t$ . The architectural depiction of the transformer encoder in the SIPFormer transformer is similar to the vision transformer (ViT) [37] in terms of hidden layer size and self-attention mechanism. However, unlike in ViT, the input to the SIPFormer transformer is not the standard positional embeddings generated by the multilayer perceptron (MLP). Instead, we generated the positional embeddings through normalized

cross correlation. Moreover, in the ViT, the linear embeddings from the image patches are generated again using MLPs, whereas in the SIPFormer transformer, the linear embeddings are generated using the SIPFormer encoder. The encodings from the positional embeddings and linear embeddings are then concatenated to pass on to the transformer encoder, as shown in Fig. 5.

In the SIPFormer transformer, the periocular scan  $x$  is first divided into nonoverlapping squared patches  $x^p \in \mathbb{R}^{(P \times P \times C_h)}$ , where  $P$  denotes the resolution of  $x^p$ , such that  $P = ((RC/n_p))^{1/2}$  and  $n_p$  denotes the number of patches. Also, each patch is cross-correlated with  $x$  to generate the positional embeddings  $x^e \in \mathbb{R}^{(P \times P \times C_h)}$ . Afterward, we obtain the flattened projections of the positional embedding  $x_i^e$  (corresponding to the patch  $x_i^p$ ), i.e.,  $f_p(x_i^e)$ , and the latent projection for the patch  $x_i^p$ , i.e.,  $l_t(x_i^p)$  through the SIPFormer encoder backbone, as shown in Figs. 4 and 5. Then, we resize  $f_p(x_i^e)$  and  $l_t(x_i^p)$  to  $k$  dimensions and compute the sequenced embeddings (for the patch  $x_i^p$ ) by summing  $l_t(x_i^p)$  with  $f_p(x_i^e)$ , i.e.,  $q_i = l_t(x_i^p) + f_p(x_i^e)$ . Moreover, repeating the same workflow for all the  $n_p$  patches yields combined projections  $q^o$ , as expressed in the following:

$$q^o = \left[ l_t(x_0^p); l_t(x_1^p); \dots; l_t(x_{(n_p-1)}^p) \right] + \left[ f_p(x_0^e); f_p(x_1^e); \dots; f_p(x_{(n_p-1)}^e) \right] \quad (1)$$

or

$$q^o = [q_0; q_1; \dots; q_{(n_p-1)}]. \quad (2)$$

The combined projections  $q^o$  are passed to the first transformer encoder, where head  $j$ ,  $q_j^o$ , is normalized to produce  $q_j^{o'}$ . Afterward,  $q_j^{o'}$  is linearly decomposed into query ( $Q_j$ ), key ( $K_j$ ), and value ( $V_j$ ) pairs via learnable weights such that  $Q_j = q_j^{o'} w_q$ ,  $K_j = q_j^{o'} w_k$ , and  $V_j = q_j^{o'} w_v$ . To compute the contextual self-attention at head  $j$ , i.e.,  $A_j$ ,  $Q_j$ , and  $K_j$  are combined via scaled dot product, their resultant scores are fused with  $V_j$ , as formulated in the following:

$$A_j(q_j^{o'}; Q_j, K_j, V_j) = \sigma \left( \frac{(Q_j K_j^T)}{\sqrt{k}} \right) V_j \quad (3)$$

where  $\sigma(\cdot)$  denotes the sigmoid function. Apart from this, the proposed contextual self-attention maps from multiple heads are concatenated together to produce contextual multihead self-attention distribution  $\varnothing_{\text{CMSA}}(q^{o'})$ , as expressed in the following:

$$\varnothing_{\text{CMSA}}(q^{o'}) = \left[ A_0(q_j^0; Q_0, K_0, V_0); A_1(q_j^1; Q_1, K_1, V_1); \dots A_{h-1}(q_j^{h-1}; Q_{h-1}, K_{h-1}, V_{h-1}) \right]. \quad (4)$$

We also want to highlight here that, in contrast to the conventional multihead attention mechanism proposed in [37] and [38], which uses softmax as an attention operator, the proposed contextual multihead self-attention scheme employs sigmoid as an operator to compute self-attention, which allows the transformer encoders to generate more vibrant attention maps without biasing itself to one particular segmentation category out of the rest within the similarly structured scans. This results in the generation of better latent projections allowing the SIPFormer decoder to accurately extract the multitrait information. Moreover, the self-attention distribution  $\varnothing_{\text{CMSA}}(q^{o'})$  is added with  $q^o$ , where the resultant embeddings are normalized and are passed to the normalized feedforward block to produce the first transformer's latent projections ( $p_{T1}$ )

$$p_{T1} = \varphi_f \left( (\varnothing_{\text{CMSA}}(q^{o'}) + q^{o'})' \right) + (\varnothing_{\text{CMSA}}(q^{o'}) + q^{o'})'. \quad (5)$$

$p_{T1}$  is passed to second transformer encoder, which produces  $p_{T2}$  in a similar manner, and  $p_{T2}$  is passed to the third transformer encoder, which produces  $p_{T3}$  projections. Since within the SIPFormer architecture, we injected three transformer encoders, so  $p_i = p_{T3}$ , and after computing  $f_d$ , they are passed to the SIPFormer decoder to segment multiocular trait representations. Table III shows the SIPFormer transformer parameters used in this study.

3) *SIPFormer Decoder*: After convolving  $f_e$  with  $p_i$ , we obtained fused feature representations  $f_d$  that are passed to the SIPFormer decoder to segment the multiocular traits. The SIPFormer decoder consists of three upsampling blocks, where each block includes the transposed C, BN, and R activations, and the skip connections (between the SIPFormer encoder and decoder). Besides, at the head of the decoder lies the softmax layer that classifies each pixel into one of the four categories (representing the background periocular, sclera, iris, and pupil regions).

Moreover, Fig. 6 shows the channel activation maps learned by the proposed SIPFormer model with and without integrating

TABLE III  
SIPFORMER TRANSFORMER PARAMETERS

Parameters	Value
Transformer Encoders	3
Encoder Layers	30
Total Layers	36
Input Image Size	96×96
Patch Size	16×16
Total Patches	36
Number of Heads	4
Projection Dimension	64
Encoder MLP Units	[64, 128]
Total Parameters	7.1 million

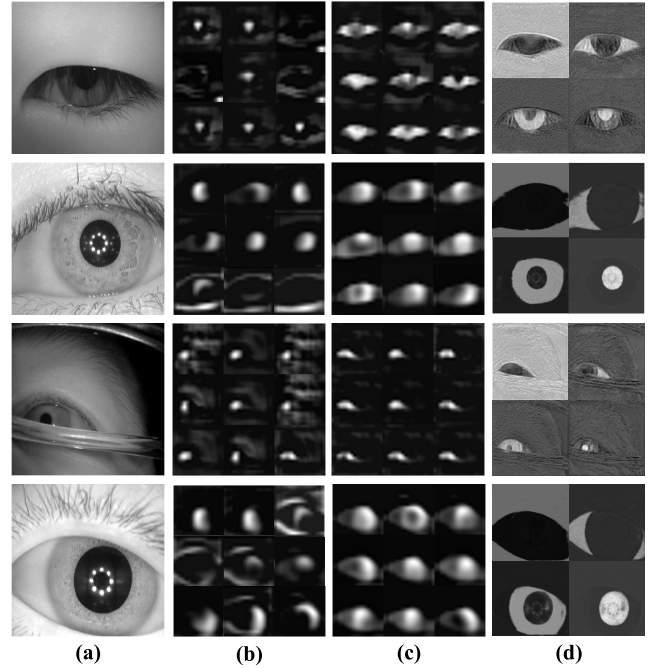


Fig. 6. Visualization of channel activation maps. (a) Input images. Nine most strongest activation channels extracted (b) without integrating SIPFormer transformer and (c) with SIPFormer transformer. (d) Class activation maps at the scorer layer of the SIPFormer decoder.

the SIPFormer transformer module. The proposed model, when combined with the SIPFormer transformer module, focuses mainly on the ocular traits in the images by neglecting the noise effects, such as eyelashes, eyelids, spectacles, and shadows, which are inherent in unconstrained acquisition settings. In contrast, when the SIPFormer transformer module is not integrated, the model is seriously affected by these noises, resulting in poor feature learning. Therefore, we can say that our proposed SIPFormer model is intrinsically robust in segmenting the multiocular traits attributed to its multihead self-attention mechanism based on the sigmoid function and the positional embeddings generated using the normalized cross correlation. Moreover, we have presented the class activation maps at the scorer layer of the SIPFormer decoder, which influenced the model to classify pixels belonging to four different classes in this study.

4) *Postprocessing*: The segmented images by the deep learning model are often noisy. As a result, we designed

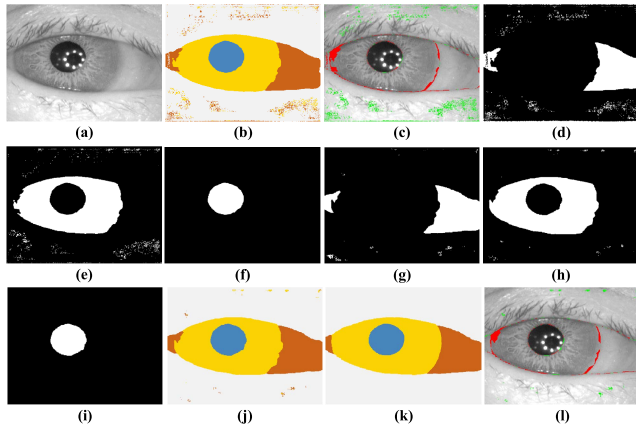


Fig. 7. Postprocessing results. (a) Original image. (b) Segmented image by the SIPFormer model. (c) Undersegmented/oversegmented pixels in (b). (d)–(f) Extract ocular masks from (b). (g)–(i) Postprocessed ocular regions mask. (j) Reconstructed segmented image after postprocessing. (k) Ground-truth image. (l) Undersegmented/oversegmented pixels in (j).

a postprocessing phase to remove stray pixels and smooth the segmented labels. For this purpose, we first extract the masks of all ocular traits (sclera, iris, and pupil) separately to perform the erosion operations using disk-shaped structural elements. Following that, we use nonlinear median filtering and 2-D convolutional blurring to minimize the noisy pixels while preserving the boundary of the ocular region. Next, we binarized these ocular masks using the thresholding technique and concatenated them to acquire a single image. The resultant image in this way may also contain some unclassified areas represented by pixels with a value of 0. Finally, we determine the locations of these empty pixels and assign the closest label (having the least Euclidean distance) to them to achieve the postprocessing images. Fig. 7 shows the illustration of the postprocessing steps to clean the segmented pixels by the proposed SIPFormer model.

5) *Training Loss Function*: The cross-entropy ( $L_c$ ) [39] or Dice loss ( $L_d$ ) [40] functions are often used to train the deep learning models for the semantic segmentation tasks [41], [42].  $L_c$  has gained popularity due to its ability to generate desirable gradients by subtracting the expected probability from the actual labels. Moreover, it greatly improves network convergence and is a suitable option for datasets with uniformly distributed classes and explicit mask annotations [43], [44], [45]. However,  $L_d$  and Tversky loss ( $L_t$ ) [46] are the preferable option for the sparse or unbalanced segmentation pixels [47]. Besides,  $L_d$  assists the model in achieving more precise segmented regions with high overlap with the corresponding ground truths (especially for cases with unbalanced classes or unclear annotations) [41], [42]. In addition,  $L_t$  offers high resilience to unbalanced classes, which further contributes to improving semantic segmentation performance [47].

Given the above, we hypothesize that synergizing two loss functions may attain the best segmentation performance. Therefore, we employed a multiobjective hybrid loss function ( $L_h$ ) in this research to improve the capacity of the proposed SIPFormer model for better recognizing the three ocular regions. We linearly combined two-tiered objective functions

( $L_d$  and  $L_t$ ) to calculate  $L_h$ , as expressed in the following:

$$L_h = \frac{1}{N} \sum_{i=1}^N (\alpha_1 L_{d,i} + \alpha_2 L_{t,i}) \quad (6)$$

$$L_{d,i} = 1 - \frac{2 \sum_{j=1}^C t_{i,j} p_{i,j}}{\sum_{j=1}^C t_{i,j}^2 + \sum_{j=1}^C p_{i,j}^2} \quad (7)$$

$$L_{t,i} = 1 - \frac{\sum_{j=1}^C t_{i,j} p_{i,j}}{\sum_{j=1}^C (t_{i,j} p_{i,j} + \beta_1 t'_{i,j} p_{i,j} + \beta_2 t_{i,j} p'_{i,j})} \quad (8)$$

where  $t_{i,j}$  shows the ground-truth labels of  $i$ th example belonging to the  $j$ th ocular class, whereas  $p_{i,j}$  denotes the predicted labels of the  $i$ th example for the  $j$ th ocular class. The terms  $t'_{i,j}$  and  $p'_{i,j}$  represent the false predicted labels, where  $t'_{i,j}$  are the ground-truth labels of the  $i$ th example belonging to the non- $j$ th class and  $p'_{i,j}$  are the predicted labels marking the  $i$ th example for the non- $j$ th class.  $N$  is the training batch size, and  $C$  specifies the total classes. The terms  $\alpha$  and  $\beta$  represent the experimentally established loss weights for achieving the best performance of the model.

### III. EXPERIMENTAL SETUP

The proposed SIPFormer framework has been implemented using the MATLAB R2022a simulation platform installed on a 64 bits Windows OS. The machine is configured as Intel Core i7-11700 @2.5 GHz, with 32-GB memory and Nvidia GeForce RTX 3090. The SIPFormer model is trained using 31 220 images, randomly chosen from the five subsets of the CASIA-IrisV4 database, as mentioned in Table II. Moreover, we augmented the data at each epoch to prevent overfitting and improve the classifier performance against the unseen data. We adopted four types of transformations to augment the data (reflection, rotation, scaling, and translation) to enable better generalization characteristics in the model.

Furthermore, the Adam optimizer [48] is employed in the proposed research to update the SIPFormer parameters during the training phase. The batch size and epochs are set to 32 and 120, respectively, allowing the network to train over 117 120 iterations with 976 iterations per epoch. Furthermore, 10 407 separate images are used for validation purposes of the SIPFormer model. We specified the validation frequency every ten epochs, enabling the SIPFormer model to validate the unseen data 12 times. In an attempt to minimize the training error on the validation set, the hyperparameters for training the SIPFormer model are determined via Bayesian optimization on 30 objective function evaluations.

### IV. SIMULATION RESULTS

In this section, we first present different ablation studies relevant to the proposed SIPFormer model. Next, we evaluated the performance of the proposed model both subjectively and objectively, as explained in the following.

#### A. Ablation Study

The ablative aspects of this research comprise: 1) determining the best weights for the  $\alpha$  and  $\beta$  parameters in the

TABLE IV

PERFORMANCE COMPARISON WITH DIFFERENT LOSS FUNCTION WEIGHTS. THE RESULTS ARE PRESENTED IN THE MEAN DSC SCORE. THE BOLD FONT REPRESENTS THE COMBINATION OF THE OPTIMAL WEIGHTS

Weight	Mean DSC Score for Different Loss Weight Sets				
	(0.3,0.7)	(0.4,0.6)	(0.5,0.5)	(0.6,0.4)	(0.7,0.3)
$(\alpha_1, \alpha_2)$	0.8871	<b>0.9023</b>	0.8923	0.8817	0.8803
$(\beta_1, \beta_2)$	0.8246	0.8492	<b>0.8746</b>	0.8576	0.8535

loss function; 2) training the network with different optimizers and loss functions to find the best combination by measuring the segmentation performance on the validation set; and 3) identifying the best backbone network for extracting features and getting the best segmentation results.

1) *Tuning the Loss Parameters:* In this experiment, we experimented with different values for loss weight parameters to determine the best combination for training the network. The  $\alpha_1$  and  $\alpha_2$  terms in (6) reflect the contribution of each loss function component toward the total loss, where  $\alpha_1$  and  $\alpha_2$  regulate the contributions of  $L_d$  [see (7)] and  $L_t$  [see (8)], respectively. Table IV shows the performance of the SIPFormer model for different combinations of the  $\alpha_1$  and  $\alpha_2$  parameters, revealing that setting  $\alpha_1 = 0.4$  and  $\alpha_2 = 0.6$  yields the best segmentation performance on the validation set with the mean Dice similarity coefficient (DSC) of 0.9023.

Similarly, the  $\beta$  factor in (8) defines the contributions of falsely predicted labels in the Traverky loss, where  $\beta_1$  and  $\beta_2$  control the contribution of false positives and false negatives. It can be observed from Table IV that the SIPFormer achieved the best segmentation results when considering the equal contribution of false positives and false negatives ( $\beta_1 = \beta_2 = 0.5$ ).

2) *Selection of Optimizer and Loss Function:* In this experiment, we trained the model using various optimizers [48], [49], [50] and loss functions [39], [40], [46] to evaluate the segmentation performance, as shown in Table V. The SIPFormer model fared best on the validation set with the ADAM +  $L_h$  configuration, achieving a mean DSC score of 0.9023. Besides, with mean DSC scores of 0.8935 and 0.9257, the ADAM +  $L_h$  arrangement achieved the best performance for segmenting the iris and pupil regions. However, the ADAM +  $L_h$  configuration obtained the second-best results for the sclera, following the best setting (SGDM +  $L_h$ ) by just 0.43%. Furthermore, compared to the second-best results (SGDM +  $L_h$ ), the mean DSC score with the ADAM +  $L_h$  configuration improves by 1.86% in ocular regions segmentation, rising from 0.8858 to 0.9023. The RMSP +  $L_c$  setup, on the other hand, exhibited the least accurate performance, with a mean DSC score of 0.7836 for segmenting the ocular regions.

3) *Selection of Backbone Network:* In this experiment, we employed different backbone networks to determine the optimal structure for extracting features and achieving the best segmentation performance on the validation set. The results are reported in Table VI, where it can be seen that the proposed SIPFormer model achieves a mean DSC score of 0.9023 for joint segmentation of the three ocular regions and exceeds the

TABLE V

PERFORMANCE COMPARISON FOR OCULAR REGIONS SEGMENTATION USING VARIOUS OPTIMIZERS AND LOSS FUNCTIONS. THE MEAN DSC SCORE METRIC IS USED TO PRESENT THE RESULTS. THE TOP PERFORMANCE IS HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST PERFORMANCE IS UNDERLINED

Optimizer + Loss	Sclera	Iris	Pupil	Mean
RMSP + $L_c$	0.7549	0.7929	0.8031	0.7836
RMSP + $L_d$	0.8016	0.8313	0.8483	0.8270
RMSP + $L_t$	0.8203	0.7746	0.8893	0.8281
RMSP + ( $L_c + L_d$ )	0.8011	0.7655	0.8383	0.8016
RMSP + $L_h$	0.8361	0.8283	0.8495	0.8379
SGDM + $L_c$	0.8053	0.7849	0.8203	0.8035
SGDM + $L_d$	0.8490	0.8701	0.8715	0.8635
SGDM + $L_t$	0.8661	0.8582	0.8758	0.8667
SGDM + ( $L_c + L_d$ )	0.8150	0.8296	0.8542	0.8329
SGDM + $L_h$	<b>0.8916</b>	0.8703	0.8957	<u>0.8858</u>
ADAM + $L_c$	0.7620	0.7654	0.8751	0.8008
ADAM + $L_d$	0.8440	0.8702	0.8832	0.8658
ADAM + $L_t$	0.8452	<u>0.8832</u>	0.8955	0.8746
ADAM + ( $L_c + L_d$ )	0.8283	0.8561	<u>0.9001</u>	0.8615
ADAM + $L_h$	<u>0.8878</u>	<b>0.8935</b>	<b>0.9257</b>	<b>0.9023</b>

Optimizers: Root mean square propagation (RMSP) [49], Stochastic gradient descent with momentum (SGDM) [50], Adaptive moment estimation (ADAM) [48].

Loss: Cross-entropy ( $L_c$ ) [39], Dice ( $L_d$ ) [40], Tversky ( $L_t$ ) [46], and Hybrid loss function ( $L_h$ ).

TABLE VI

PERFORMANCE COMPARISON FOR OCULAR REGIONS SEGMENTATION USING DIFFERENT BACKBONES STRUCTURES. THE MEAN DSC SCORE METRIC IS USED TO PRESENT THE RESULTS. THE TOP PERFORMANCE IS HIGHLIGHTED IN BOLD, WHILE THE SECOND-BEST PERFORMANCE IS UNDERLINED

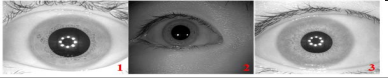

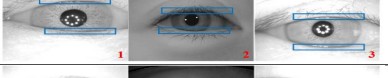
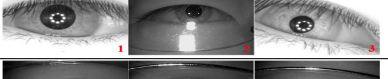
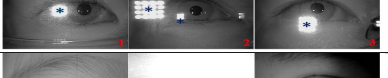

Backbone	Sclera	Iris	Pupil	Mean
RegNet [51]	0.8744	0.8254	0.8875	0.8624
EfficientNetV2 [52]	0.8875	0.8341	<u>0.8996</u>	0.8737
ResNet-101 [53]	<b>0.8945</b>	<u>0.8730</u>	0.8552	<u>0.8742</u>
SIPFormer	<u>0.8878</u>	<b>0.8935</b>	<b>0.9257</b>	<b>0.9023</b>

second-best results by 3.21%. Moreover, the SIPFormer model produced 2.35% and 2.90% better performance for segmenting the iris and pupil regions, whereas for segmenting sclera, the proposed framework achieves the second-best results with a mean DSC score of 0.8878, lagging the best results by 0.75%.

4) *Effects of Image Resolution and Patch Size:* In this experiment, we studied the effect of image resolution and patch size (PS) for the proposed SIPFormer model on the validation set. We tested our model performance based on three image resolutions ( $320 \times 512$ ,  $576 \times 768$ , and  $640 \times 832$ ) and three PSs ( $8 \times 8$ ,  $16 \times 16$ , and  $32 \times 32$ ), as shown in Fig. 8. Furthermore, the performance of the model is analyzed in terms of two parameters: segmentation accuracy using the mean Dice score and model inference time in terms of frames per second (FPS). From the results in Fig. 8, we can see that increasing the image resolution and the number of patches improves the segmentation accuracy slightly [see Fig. 8(a)]. However, the SIPFormer model efficiency is inversely proportional to the image resolution and PS. Thus, variants with smaller PSs and higher image resolution are computationally far more expensive [see Fig. 8(b)]. Therefore,



TABLE VII  
CATEGORIES OF OCULAR OCCLUSIONS FOR SUBJECTIVE EVALUATION

Category	Description	Sample Images
Clear	This category includes images that do not have any ocular obstructions. The sclera, iris, and pupil regions are clearly apparent in these photos, as seen in the sample images.	
Eyelashes Occlusion	This category refers to obstruction in the ocular region, predominantly by the eyelashes, as illustrated in the sample images. The yellow bounding box represents the eyelash occlusion.	
Eyelids Occlusion	The ocular regions in these images are partially occluded by the upper, lower, or both eyelids. Blue bounding boxes highlight the occluded ocular regions.	
Heavily Occluded	We categorized the images as heavily occluded when the ocular region is obscured by several entities such as eyelashes, eyelids, and spectacles reflections.	
Spectacle Reflection	This category refers to images with spectacle reflections obstructing the ocular region. These reflections can occur for various reasons, including camera flash, light, etc. The spectacle reflections in the sample images are marked by an asterisk (*).	
Truncated	The ocular region in this category is truncated and not completely visible due to poor acquisition of images.	

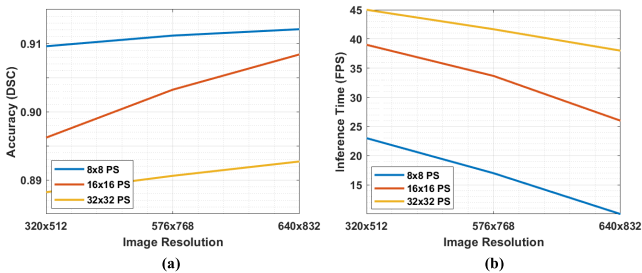


Fig. 8. Effects of image resolution and PS on (a) segmentation accuracy and (b) model inference time.

we opted for moderate settings ( $576 \times 768$  image resolution with  $16 \times 16$  PS) in the proposed research.

### B. Subjective Evaluation

Periocular components, such as eyelashes, eyelids, and spectacle reflection, often obstruct the ocular region. Therefore, we analyzed the performance of our trained SIPFormer model subjectively against the various occlusion categories, as shown in Table VII. To conduct this experiment, we randomly selected 200 images for each occlusion category from the testing dataset.

Fig. 9 compares the segmentation results of the SIPFormer with other state-of-the-art methods [37], [54], [55], [56], [57], [58]. To demonstrate the results, we randomly selected 200 images from each occlusion category, as mentioned in Table VII. Here, we can observe that the segmented pixels by the proposed SIPFormer model [see Fig. 9(j)] are more precise in general compared to other methods [see Fig. 9(c)–(i)], producing a lesser number of false positive and false negative pixels, as highlighted with the green and red, respectively. Moreover, the first two rows in Fig. 9 show clear images, which do not contain occlusion in the sclera, iris, and pupil regions. The SIPFormer model achieved the best results for images in this category, where the segmented regions are nearly identical to the corresponding ground truths [see Fig. 9(b)]. The higher accuracy in such cases is perhaps due to the discernible and clear contours of each ocular component.

The following two rows present the segmentation results for images containing obstructions in the ocular regions due to eyelashes. The proposed framework showed a good generalization for such images and classified the majority of ocular pixels correctly, with some false negatives (undersegmentation). Similarly, the proposed SIPFormer model produced promising results for images predominated with eyelids occlusion, which mainly truncates the contour and symmetry of the iris region. The SIPFormer model preserved the symmetry of the ocular regions as in the ground truths for these images by precluding the obstructed sclera and iris regions, as evident from the segmentation results.

Furthermore, we evaluated the segmentation performance for images in the heavily occluded category. The ocular region in this category is occluded through multiple obstructions such as eyelashes, eyelids, and reflections. Compared with the other categories, the segmentation performance of SIPFormer for this set is not as accurate, producing some false positive and false negative pixels, as evident from Fig. 9. Moreover, we validated the performance of the SIPFormer model for images containing spectacle reflections, as shown in the ninth and tenth row of Fig. 9. The proposed model generally produced good results for such images. However, it segmented some pixels around the contours of ocular regions as false positives and negatives.

Finally, we computed the segmentation accuracy of SIPFormer for poorly acquired images having truncated ocular regions, as shown in the final two rows of Fig. 9. We categorized images as poorly acquired if: 1) the ocular region is not entirely visible due to spectacle frame or poorly aligned shooting angle and 2) the ocular region is partially or wholly dark due to shadow or illumination differences. Generally, the SIPFormer demonstrated superior segmentation accuracy for such cases with some false negative pixels for the iris and sclera regions. The prime reason for these false negatives is the incomplete and irreconcilable contours of the iris and sclera regions in this category. To summarize, the segmented pixels by the SIPFormer model for pupil class overlap well with the corresponding ground-truth pixels. However, the segmented

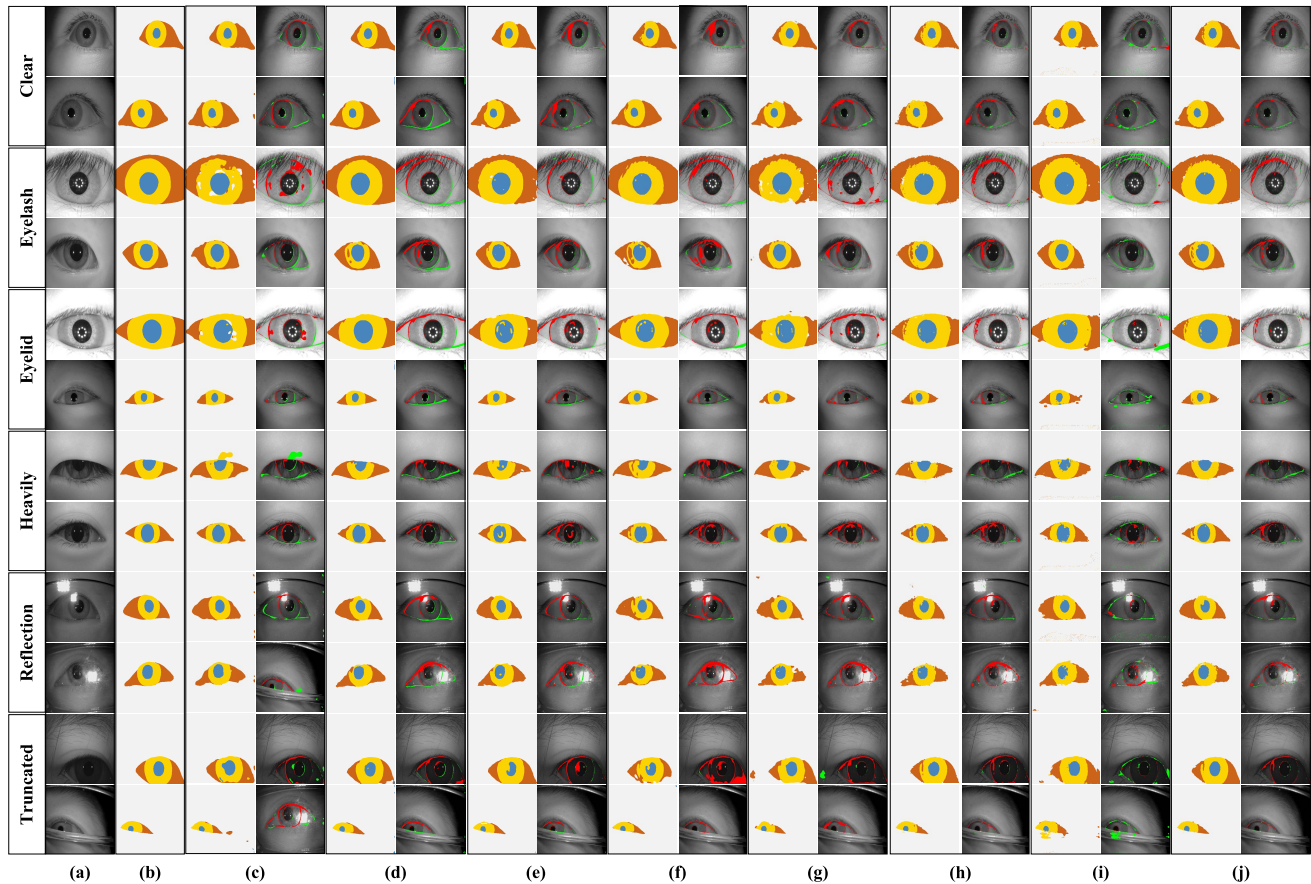


Fig. 9. Performance comparison for ocular regions segmentation. (a) Original images. (b) Ground-truth labels. (c) AUnet. (d) RefineNet. (e) Segnet. (f) SegFormer. (g) FCN-Resnet101. (h) ViT + SIPFormer ED. (i) DeepLabv3+. (j) SIPFormer.

pixels near the sclera–iris boundary are slightly less accurate. Next, we measure the exact overlap per class for each of the above-discussed occlusion categories.

### C. Objective Evaluation

In this section, we present the objective evaluation of the proposed SIPFormer framework for jointly segmenting the sclera, iris, and pupil. We objectively evaluated the performance of SIPFormer via three main experiments, as explained in the following.

1) *Evaluation on Test Datasets*: In this experiment, we first present the performance of the SIPFormer model against all of the five CASIA-IrisV4 test sets using different metrics. The results are reported in Table VIII, where it can be observed that our proposed framework generally performed well for all five CASIA-IrisV4 subsets, achieving the mean intersection over union (IoU) and DSC scores of 0.8226 and 0.9025, respectively. Also, we can analyze that the SIPFormer performed slightly better on the CIS and CIL datasets compared to the other three datasets (CII, CITW, and CIT). This is perhaps because the CII dataset contains close-up eye images, while all the other datasets contain images captured from a distance. Also, the ratio of CII images is low compared to other datasets.

In addition, the CIT dataset comprises the most complex images with more intraclass variances, including spectacles and specular reflections. Moreover, the CIT dataset with 20 000

TABLE VIII  
DATASETWISE SEGMENTATION PERFORMANCE OF THE SIPFORMER MODEL USING VARIOUS METRICS. THE BOLD AND UNDERLINED VALUES REPRESENT THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY

Metrics	Datasets (All Classes)					
	CII	CIS	CIL	CITW	CIT	Mean
Precision	0.8776	<b>0.8978</b>	0.883	0.8676	<u>0.8927</u>	0.8837
Recall	0.9216	<u>0.9322</u>	<b>0.9357</b>	0.9273	0.9244	0.9282
Specificity	0.9566	<b>0.9640</b>	0.9583	0.9519	<u>0.9623</u>	0.9586
IoU	0.8172	<b>0.8432</b>	<u>0.8329</u>	0.7874	0.8322	0.8226
DSC	0.8988	<b>0.9143</b>	<u>0.9084</u>	0.8829	0.9079	0.9025
Nice1	0.0522	<b>0.0439</b>	0.0474	0.0542	<u>0.0471</u>	0.0490
Nice2	0.0609	<b>0.0519</b>	<u>0.0530</u>	0.0604	0.0566	0.0566

images is the biggest, and achieving the best precision for such a huge dataset is challenging. In addition, the images in the CITW dataset were shot during the annual festival in Beijing, and it is the only dataset where images were captured in an outdoor setting. This dataset was distinct from other datasets due to the variation in lighting, which is why our proposed model obtained the lowest accuracy for CITW. However, it is essential to note that despite these differences in the datasets, the segmentation results for each dataset change by a small margin, which affirms the resilience and reliability of the SIPFormer framework regardless of the datasets.

TABLE IX

CLASSWISE SEGMENTATION PERFORMANCE OF THE SIPFORMER MODEL USING VARIOUS METRICS. THE BOLD AND UNDERLINED VALUES REPRESENT THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY

Metrics	Classes (All Datasets)				
	Periocular	Sclera	Iris	Pupil	Mean
Precision	0.8234	0.8767	<u>0.9126</u>	<b>0.9223</b>	0.8837
Recall	0.9168	<u>0.9288</u>	0.9228	<b>0.9446</b>	0.9282
Specificity	0.9344	0.9563	<u>0.9705</u>	<b>0.9734</b>	0.9586
IoU	0.7662	0.8214	<u>0.8478</u>	<b>0.8549</b>	0.8226
DSC	0.8675	0.9018	<u>0.9176</u>	<b>0.9229</b>	0.9025
Nice1	0.0700	0.0506	<u>0.0415</u>	<b>0.0338</b>	0.0490
Nice2	0.0744	0.0575	<u>0.0534</u>	<b>0.0410</b>	0.0566

Moreover, we computed the classwise performance of the proposed model on the complete test data containing 10407 images, as shown in Table IX. Considering the classwise performance, we can see that the SIPFormer model performed best for pupil class followed by iris, as shown in Table IX. The higher IoU and DSC scores for both pupil and iris classes represent the higher overlapped region and similarity index between the ground-truth and predicted labels of these classes. In contrast, the same metrics gave relatively low scores for the periocular and sclera classes. This can be because the CASIA-IrisV4 datasets are in grayscale. The pixel intensities in the periocular and sclera region are relatively close and are not distinguishable from the other two classes (iris and pupil). As a result, the network performance slightly declined in predicting the periocular and sclera labels. Similarly, the higher values of Nice1 and Nice2 metrics for both periocular and sclera classes reflect the higher ratio of false predicted labels for these classes compared to iris and pupil.

Furthermore, we analyzed the performance of the SIPFormer using the receiver operating characteristic (ROC) and precision–recall curves of each class, as shown in Fig. 10. We generated the ROC and precision–recall curves for each class by varying the pixel classification threshold between 0 and 1 with a step size of 0.005. It can be observed from the ROC curves in Fig. 10(a) that our proposed SIPFormer model is trained skillfully in correctly predicting the positive labels in each class. The higher area under the curve (AUC) value for each class also quantifies the superior predictive performance of the SIPFormer model. Since, in our study, there is a considerable imbalance between the classes. Therefore, precision–recall curves [Fig. 10(b)] show the advantages of the SIPFormer algorithm more intuitively due to the absence of true negatives in precision and recall equations. The higher mean average precision (mAP) values for each class show that the SIPFormer is trained precisely, returns accurately (high precision), and correctly predicts the most positive results (high recall).

### 2) Evaluation Based on Ocular Occlusion Categories:

In this experiment, we evaluated the performance of the SIPFormer model for different kinds of occlusion using various metrics. We used 200 randomly selected images from each occlusion category, as defined in Table VII. Fig. 9 and Table X show that the SIPFormer worked best with clear

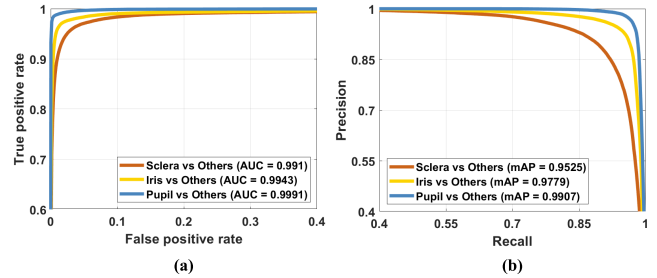


Fig. 10. (a) ROC curve and (b) precision–recall curve for each class.

TABLE X

IMAGES CATEGORYWISE SEGMENTATION PERFORMANCE OF THE SIPFORMER MODEL USING VARIOUS METRICS. THE BOLD AND UNDERLINED VALUES REPRESENT THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY

Category	DSC	IoU	Recall	Specific	Precision	Nice2	Nice1
Clear	<b>0.9382</b>	<b>0.8820</b>	<b>0.9662</b>	<u>0.9811</u>	<b>0.9301</b>	<b>0.0220</b>	<b>0.0230</b>
Eyelash	0.9083	0.8303	<u>0.9628</u>	<b>0.9933</b>	0.8890	<u>0.0380</u>	<u>0.0393</u>
Eyelid	<u>0.9289</u>	<u>0.8657</u>	0.9245	0.9688	<u>0.9137</u>	0.0533	0.0536
Heavy	0.8490	0.7361	0.8759	0.9056	0.8117	0.1093	0.0821
Reflection	0.8589	0.7511	0.8784	0.9168	0.8248	0.1024	0.0820
Truncated	0.8970	0.8117	0.9375	0.9578	0.8591	0.0407	0.0409
Mean	0.8967	0.8128	0.9242	0.9539	0.8714	0.0609	0.0535

category images, followed by those with eyelid and eyelash obstruction, whereas its performance relatively declined for images in the truncated and reflection categories. Furthermore, the proposed framework performed the least accurately with heavily obscured images. The latter three categories can be termed challenging cases in our study. In addition, fewer images belong to these categories in the whole dataset. Consequently, this lack of data may have resulted in insufficient network training to manage such exceptions properly.

3) *Comparison With State-of-the-Art Literature:* In this experiment, we evaluated the segmentation accuracy of the SIPFormer model to other state-of-the-art models, as shown in Table XI. With a mean DSC score of 0.9018, the proposed framework provides the best segmentation results for the sclera class, exceeding the second-best results by 0.92%. In contrast, it demonstrated the second-best performance in segmenting the iris and the third-best performance in segmenting the pupil, trailing the best results by 0.97% and 1.12%, respectively. In addition, the proposed SIPFormer framework outperforms the second-best method for segmentation of all ocular classes by 0.26%.

Furthermore, we have evaluated the computational complexity of the proposed framework with other state-of-the-art methods using the consistent simulation environment, as shown in Table XII. All the models in Table XII have been trained and evaluated using the same system and hardware specifications. The computational complexity of the models is evaluated using an Intel Core i7-11700 CPU @2.5 GHz, with 32-GB memory and Nvidia GeForce RTX 3090. Moreover, the training hyperparameters for all models in Table XII are adjusted using the Bayesian optimization on 30 objective function evaluations to ensure the best training performances. Here, it can be observed that with 37.2 million parameters, the SIPFormer framework requires 92.13% fewer parameters

TABLE XI

SEGMENTATION PERFORMANCE COMPARISON WITH THE STATE OF THE ART USING MEAN DSC SCORE. THE BOLD AND UNDERLINED VALUES REPRESENT THE BEST AND SECOND-BEST RESULTS, RESPECTIVELY

Model (Backbone)	Test DSC scores			
	Sclera	Iris	Pupil	Combined
AUnet [54]	0.8482	0.8574	0.8819	0.8625
RefineNet (ResNet-101) [55]	0.8374	0.8852	0.9033	0.8753
SegNet (VGG-16) [56]	0.8750	0.8871	0.8971	0.8864
SegFormer (MiT-B3) [57]	0.8503	0.8895	0.9314	0.8904
FCN (ResNet-101)	0.8654	0.8985	<b>0.9332</b>	0.8990
ViT [37] + SIPFormer ED	<u>0.8936</u>	0.9067	<u>0.9320</u>	0.9108
DeepLabv3+ (Inception-ResNet-v2) [58]	0.8815	<b>0.9265</b>	0.9272	<u>0.9117</u>
SIP Former	<b>0.9018</b>	<u>0.9176</u>	0.9229	<b>0.9141</b>

TABLE XII

COMPUTATIONAL COMPLEXITY ANALYSIS WITH THE STATE OF THE ART USING NVIDIA GTX 3090 GPU. THE BOLD AND UNDERLINED VALUES REPRESENT THE BEST AND SECOND-BEST PERFORMANCES, RESPECTIVELY

Model (Backbone)	Computational Complexity			
	Resolution	Params (M)	Runtime (ms)	FPS
AUnet [54]	288×288	132.7	48.14	19
RefineNet (ResNet-101) [55]	299×299	117.5	40.13	25
SegNet (VGG-16) [56]	299×299	<b>29.4</b>	30.91	32
SegFormer (MiT-B3) [57]	288×288	45.2	33.73	29
FCN (ResNet-101)	299×299	44.6	33.07	30
ViT [37] + SIPFormer ED	576×768	38.4	<u>27.25</u>	<u>35</u>
DeepLabv3+ (Inception-ResNet-v2) [58]	299×299	71.1	38.52	26
SIP Former	576×768	<u>37.2</u>	<b>26.70</b>	<b>37</b>

than the second-best performing method [58]. Besides, it can process 37 FPS and requires 26.70 milliseconds only to process a single image, making it suitable for real-world biometric applications.

## V. CONCLUSION AND DISCUSSION

The motivation of this work is to segment multiple ocular traits simultaneously toward devising a multimodal ocular segmentation framework. For this purpose, we employed a novel framework called SIPFormer, which fuses the transformer projections with the deep features at the encoder side to boost the separation between interclass distributions. Besides, the proposed SIPFormer model obviates the need for high computational resources due to its lesser number of parameters (around 30 million), unlike in other popular semantic segmentation architectures such as RefineNet [55], SegNet [56], and DeepLabv3+ [58].

Moreover, for several images in the datasets, we observed the difference between the black intensities in the ocular region to be very small and clustered predominantly around the lower or middle range. Therefore, in this work, we enhanced the ocular traits using preprocessing stage and subsequently suppressed the information on the periocular components. In addition, we removed various reflections in the pristine images via an improved holes filling strategy to achieve the preprocessed scans, as shown in Fig. 3. Furthermore, the

classes in the training data are highly imbalanced, with the periocular being the dominant class having over 75% of the labels. On the contrary, the pupil class had the least labels (less than 5%). This kind of bias can lead the network to ignore marginalized classes. Therefore, we used the inverse frequency weighting approach to balance the classes by assigning increased weights to the underrepresented classes (sclera, iris, and pupil). Also, we used the hybrid loss function to account for both the Dice and Tversky losses, enabling the model to train more precisely and converge quickly.

Furthermore, we conducted several experiments to evaluate the performance of the SIPFormer framework. We computed the segmentation accuracy of each class for various occlusion categories, as shown in Fig. 9 and Table X. Moreover, the proposed framework is tested over multiple challenging datasets using various evaluation metrics, as shown in Table VIII. Besides, we present the comparative analysis between the proposed SIPFormer and existing state-of-the-art algorithms, as shown in Tables XI and XII. We have also compared the performance of the proposed SIPFormer model with the standard ViT [37] + SIPFormer ED unit. We trained the ViT (with three transformer encoders) from scratch on the CASIA datasets, and the segmentation results are reported in Table XI. Here, we can observe that the SIPFormer achieved higher segmentation accuracy compared to the ViT variant. This is perhaps because the positional embeddings in SIPFormer, unlike in standard ViT, are generated through normalized cross correlation between the original image and image patches. Furthermore, the latent projections of the decomposed image are obtained through the SIPFormer encoder and combined with the flattened positional encodings to feed to the SIPFormer transformer. It provides the SIPFormer model with better and more robust feature learning capability resulting in high accuracy in segmenting the multiocular traits in the proposed study.

The simulation results in this research demonstrate the optimal performance of the SIPFormer framework in segmenting the multiocular biometric traits. The proposed model can be refined in the future using more publicly available datasets with more challenging and unique ocular images (both from near and far). Also, the segmented results from the SIPFormer model can be used for implementing a multiocular biometric recognition system.

## ACKNOWLEDGMENT

The authors are grateful to the Chinese Academy of Sciences Institute of Automation for making their datasets publicly available for research purposes. They have used five subsets of the CASIA-IrisV4 database to perform this study [36].

## REFERENCES

- [1] J. M. Smereka, "A new method of pupil identification," *IEEE Potentials*, vol. 29, no. 2, pp. 15–20, Mar. 2010.
- [2] P. Corcoran and C. Costache, "Smartphones, biometrics, and a brave new world," *IEEE Technol. Soc. Mag.*, vol. 35, no. 3, pp. 59–66, Sep. 2016.
- [3] S. Thavalengal and P. Corcoran, "User authentication on smartphones: Focusing on iris biometrics," *IEEE Consum. Electron. Mag.*, vol. 5, no. 2, pp. 87–93, Apr. 2016.

- [4] S. Bazrafkan, S. Thavalengal, and P. Corcoran, "An end to end deep neural network for iris segmentation in unconstrained scenarios," *Neural Netw.*, vol. 106, pp. 79–95, Oct. 2018.
- [5] A. S. Al-Waisy, R. Qahwaji, S. Ipson, S. Al-Fahdawi, and T. A. M. Nagem, "A multi-biometric iris recognition system based on a deep learning approach," *Pattern Anal. Appl.*, vol. 21, no. 3, pp. 783–802, Aug. 2018.
- [6] B. Hassan, R. Ahmed, B. Li, O. Hassan, and T. Hassan, "Autonomous framework for person identification by analyzing vocal sounds and speech patterns," in *Proc. 5th Int. Conf. Control, Autom. Robot. (ICCAR)*, Apr. 2019, pp. 649–653.
- [7] F. N. Sibai, H. I. Hosani, R. M. Naqbi, S. Dhanhani, and S. Shehhi, "Iris recognition using artificial neural networks," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 5940–5946, May 2011.
- [8] Z. Zhou, Y. Du, N. Thomas, and E. Delp, "A new human identification method: Sclera recognition," *IEEE Trans. Syst., Man, Cybern., A, Syst., Hum.*, vol. 42, no. 3, pp. 571–583, May 2012.
- [9] S.-Y. He and C.-P. Fan, "SIFT features and SVM learning based sclera recognition method with efficient sclera segmentation for identity identification," in *Proc. IEEE Int. Conf. Artif. Intell. Circuits Syst. (AICAS)*, Mar. 2019, pp. 297–298.
- [10] M. S. Maheshan, B. S. Harish, and N. Nagadarshan, "A convolution neural network engine for sclera recognition," *Int. J. Interact. Multimedia Artif. Intell.*, vol. 6, no. 1, p. 78, 2020.
- [11] R. Mesbah, B. McCane, and S. Mills, "Conditional random fields incorporate convolutional neural networks for human eye sclera semantic segmentation," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 768–773.
- [12] R. Hentati, M. Hentati, and M. Abid, "Development a new algorithm for iris biometric recognition," *Int. J. Comput. Commun. Eng.*, vol. 1, no. 3, p. 283, 2012.
- [13] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Iris recognition after death," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1501–1514, Jun. 2019.
- [14] N. Susitha and R. Subban, "Reliable pupil detection and iris segmentation algorithm based on SPS," *Cognit. Syst. Res.*, vol. 57, pp. 78–84, Oct. 2019.
- [15] N. Nugrahaningsih and M. Porta, "Pupil size as a biometric trait," in *Proc. Int. Workshop Biometric Authentication*. Cham, Switzerland: Springer, 2014, pp. 222–233.
- [16] B. Hassan, S. Qin, T. Hassan, R. Ahmed, and N. Werghi, "Joint segmentation and quantification of chorioretinal biomarkers in optical coherence tomography scans: A deep learning approach," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–17, 2021.
- [17] R. Ahmed, Y. Chen, B. Hassan, L. Du, T. Hassan, and J. Dias, "Hybrid machine-learning-based spectrum sensing and allocation with adaptive congestion-aware modeling in CR-assisted IoV networks," *IEEE Internet Things J.*, vol. 9, no. 24, pp. 25100–25116, Dec. 2022.
- [18] A. Khan, B. Hassan, S. Khan, R. Ahmed, and A. Abuassba, "DeepFire: A novel dataset and deep transfer learning benchmark for forest fire detection," *Mobile Inf. Syst.*, vol. 2022, pp. 1–14, Apr. 2022.
- [19] R. Ahmed, Y. Chen, and B. Hassan, "Deep residual learning-based cognitive model for detection and classification of transmitted signal patterns in 5G smart city networks," *Digit. Signal Process.*, vol. 120, Jan. 2022, Art. no. 103290.
- [20] L. A. Zanlorensi, R. Laroca, E. Luz, A. S. Britto, L. S. Oliveira, and D. Menotti, "Ocular recognition databases and competitions: A survey," *Artif. Intell. Rev.*, vol. 55, no. 1, pp. 129–180, Jan. 2022.
- [21] D. Zhu, J. Li, H. Li, J. Peng, X. Wang, and X. Zhang, "A less-constrained sclera recognition method based on Stem-and-leaf branches network," *Pattern Recognit. Lett.*, vol. 145, pp. 43–49, May 2021.
- [22] G. Sahin and O. Susuz, "Encoder–decoder convolutional neural network based iris-sclera segmentation," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.
- [23] R. A. Naqvi and W.-K. Loh, "Sclera-Net: Accurate sclera segmentation in various sensor images based on residual encoder and decoder network," *IEEE Access*, vol. 7, pp. 98208–98227, 2019.
- [24] J. Daugman, "Information theory and the iricode," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 2, pp. 400–409, Feb. 2016.
- [25] W. Dong, Z. Sun, and T. Tan, "Iris matching based on personalized weight map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1744–1757, Sep. 2011.
- [26] J. K. Pillai, V. M. Patel, R. Chellappa, and N. K. Ratha, "Secure and robust iris recognition using random projections and sparse representations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1877–1893, Sep. 2011.
- [27] R. R. Jha, G. Jaswal, D. Gupta, S. Saini, and A. Nigam, "PixISegNet: Pixel-level iris segmentation network using convolutional encoder–decoder with stacked hourglass bottleneck," *IET Biometrics*, vol. 9, no. 1, pp. 11–24, Jan. 2020.
- [28] K. Nguyen, C. Fookes, A. Ross, and S. Sridharan, "Iris recognition with off-the-shelf CNN features: A deep learning perspective," *IEEE Access*, vol. 6, pp. 18848–18855, 2017.
- [29] T. Zhao, Y. Liu, G. Huo, and X. Zhu, "A deep learning iris recognition method based on capsule network architecture," *IEEE Access*, vol. 7, pp. 49691–49701, 2019.
- [30] Y.-H. Yiu et al., "DeepVOG: Open-source pupil segmentation and gaze estimation in neuroscience using deep learning," *J. Neurosci. Methods*, vol. 324, Aug. 2019, Art. no. 108307.
- [31] K. Kitazumi and A. Nakazawa, "Robust pupil segmentation and center detection from visible light images using convolutional neural network," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2018, pp. 862–868.
- [32] F. Vera-Olmos, E. Pardo, H. Melero, and N. Malpica, "DeepEye: Deep convolutional network for pupil detection in real environments," *Integr. Comput.-Aided Eng.*, vol. 26, no. 1, pp. 85–95, 2019.
- [33] A. J.-W. Whang et al., "Pupil size prediction techniques based on convolution neural network," *Sensors*, vol. 21, no. 15, p. 4965, 2021.
- [34] L. Shi, C. Wang, F. Tian, and H. Jia, "An integrated neural network model for pupil detection and tracking," *Soft Comput.*, vol. 25, no. 15, pp. 10117–10127, Aug. 2021.
- [35] W.-L. Ou, T.-L. Kuo, C.-C. Chang, and C.-P. Fan, "Deep-learning-based pupil center detection and tracking technology for visible-light wearable gaze tracking devices," *Appl. Sci.*, vol. 11, no. 2, p. 851, Jan. 2021.
- [36] *CASIA Iris Image Database V4.0*. Accessed: Mar. 17, 2022. [Online]. Available: <http://biometrics.idealtest.org/>
- [37] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [38] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.
- [39] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: A Monte Carlo cross-entropy approach," *Bioinformatics*, vol. 23, no. 13, pp. 1607–1615, Jul. 2007.
- [40] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.
- [41] B. Hassan, S. Qin, T. Hassan, M. U. Akram, R. Ahmed, and N. Werghi, "CDC-Net: Cascaded decoupled convolutional network for lesion-assisted detection and grading of retinopathy using optical coherence tomography (OCT) scans," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103030.
- [42] B. Hassan et al., "Deep learning based joint segmentation and characterization of multi-class retinal fluid lesions on OCT scans for clinical use in anti-VEGF therapy," *Comput. Biol. Med.*, vol. 136, Sep. 2021, Art. no. 104727.
- [43] R. Ahmed, Y. Chen, and B. Hassan, "Deep learning-driven opportunistic spectrum access (OSA) framework for cognitive 5G and beyond 5G (B5G) networks," *Ad Hoc Netw.*, vol. 123, Dec. 2021, Art. no. 102632.
- [44] A. Khan, S. Khan, B. Hassan, and Z. Zheng, "CNN-based smoker classification and detection in smart city application," *Sensors*, vol. 22, no. 3, p. 892, Jan. 2022.
- [45] T. Hassan, B. Hassan, M. U. Akram, S. Hashmi, A. H. Taguri, and N. Werghi, "Incremental cross-domain adaptation for robust retinopathy screening via Bayesian deep learning," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.
- [46] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," in *Proc. Int. Workshop Mach. Learn. Med. Imag. Cham, Switzerland: Springer*, 2017, pp. 379–387.
- [47] T. Hassan, B. Hassan, A. ElBaz, and N. Werghi, "A dilated residual hierarchically fashioned segmentation framework for extracting Gleason tissues and grading prostate cancer from whole slide images," in *Proc. IEEE Sensors Appl. Symp. (SAS)*, Aug. 2021, pp. 1–6.
- [48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [49] G. Hinton, N. Srivastava, and K. Swersky, "Neural networks for machine learning," *Coursera, Video Lectures*, vol. 264, no. 1, pp. 2146–2153, 2012.
- [50] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: MIT Press, 2012.

- [51] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar, "Designing network design spaces," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10428–10436.
- [52] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
- [53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [54] S. Mazhar et al., "AUnet: A deep learning framework for surface water channel mapping using large-coverage remote sensing images and sparse scribble annotations from OSM data," *Remote Sens.*, vol. 14, no. 14, p. 3283, 2022.
- [55] G. Lin, F. Liu, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for dense prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1228–1242, May 2020.
- [56] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Jan. 2017.
- [57] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.
- [58] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2018, pp. 801–818.



**Bilal Hassan** received the Ph.D. degree in pattern recognition and intelligent systems from Beihang University, Beijing, China, in 2022.

He is currently a Post-Doctoral Fellow with the Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His research interests lie in the domain of computer vision, medical imaging, wireless communication, and control systems.



**Taimur Hassan** (Member, IEEE) received the Ph.D. degree in computer engineering from the National University of Sciences and Technology, Islamabad, Pakistan, in 2019.

He is currently working as a Post-Doctoral Fellow with the Khalifa University Center for Autonomous Robotic Systems (KUCARS) and the Center for Cyber-Physical Systems (C2PS), Department of Electrical Engineering and Computer Science, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. He has led many

local and foreign research projects. His research interests lie in the fields of medical imaging, computer vision, deep learning, the Internet of Things, and robotics.

Dr. Hassan's Ph.D. research won the Gold Award in the Research and Development Category at the Pakistan Software Houses Association for IT and ITes (P@SHA) ICT Awards in 2016, the Gold Award in the Research and Development Category at the Asia Pacific ICT Alliance (APICTA) Awards in 2016, and the Gold Award in the Artificial Intelligence category at P@SHA ICT Awards in 2018. He was a recipient of many national and international awards.



**Ramsha Ahmed** received the B.E. degree in telecommunication engineering and the M.S. degree in information security from the National University of Sciences and Technology, Islamabad, Pakistan, in 2013 and 2017, respectively, and the Ph.D. degree in information and communication engineering from the University of Science and Technology Beijing, Beijing, China, in 2022.

She is currently a Post-Doctoral Fellow with the Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. Her research interests include medical imaging, wireless communication, information security, the Internet of Things (IoT), and machine vision-related applications.



**Naoufel Werghi** (Senior Member, IEEE) received the Ph.D. degree in computer vision from the University of Strasbourg, Strasbourg, France, in 1996.

He has been a Research Fellow with the Division of Informatics, The University of Edinburgh, Edinburgh, U.K., and a Lecturer with the Department of Computer Sciences, University of Glasgow, Glasgow, U.K. He has also been a Visiting Professor with the Department of Electrical and Computer Engineering, University of Louisville, Louisville, KY, USA. He is currently an Associate Professor

with the Department of Electrical and Computer Engineering, Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates. His main area of research is image analysis and interpretation, where he has been leading several funded projects in the areas of biometrics, medical imaging, and intelligent systems.



**Jorge Dias** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Coimbra, Coimbra, Portugal, in 1994.

He coordinated the Artificial Perception Group, Institute of Systems and Robotics, University of Coimbra. He is currently a Full Professor with the Khalifa University of Science and Technology, Abu Dhabi, United Arab Emirates, where he is also the Deputy Director of the Center of Autonomous Robotic Systems. His expertise is in the area of artificial perception (computer vision and robotic

vision) and has contributions on the field since 1984. He has been a principal investigator and a consortia coordinator from several research international projects and coordinates the research group on computer vision and artificial perception from the Khalifa University Center for Autonomous Robotic Systems (KUCARS). He has published several articles in the area of computer vision and robotics that include more than 300 publications in international journals and conference proceedings and recently published book on probabilistic robot perception that addresses the use of statistical modeling and artificial intelligence for perception, planning, and decision in robots. He was the Project Coordinator of two European Consortium for the Projects "Social Robot" and "GrowMeUP" that were developed to support the inclusivity and wellbeing for of the elderly generation.